Transcriptome assembly for non-model Apinae bees: reference or *de novo* approach?

Araujo, Natália de Souza & **Arias**, Maria Cristina
Institute of Bioscience
University of São Paulo - Brazil

Introduction

RNA-Seq is a cost-effective method to characterize the gene set of species under interest. Nevertheless, the reconstruction of all full-length transcripts based on short nucleotide reads represents a substantial computational challenge.

There are two main strategies for transcriptome assembly, the Mapping-first and the Assembly-first (*de novo*). The first one is based on the alignment of all reads to a reference genome. It is less computationally intensive and, in principle, provides maximum sensitivity. However it demands an accurate mapped genome as reference. Conversely, the Assembly-first strategy assemble the reads in contigs not using a reference genome as a guide.

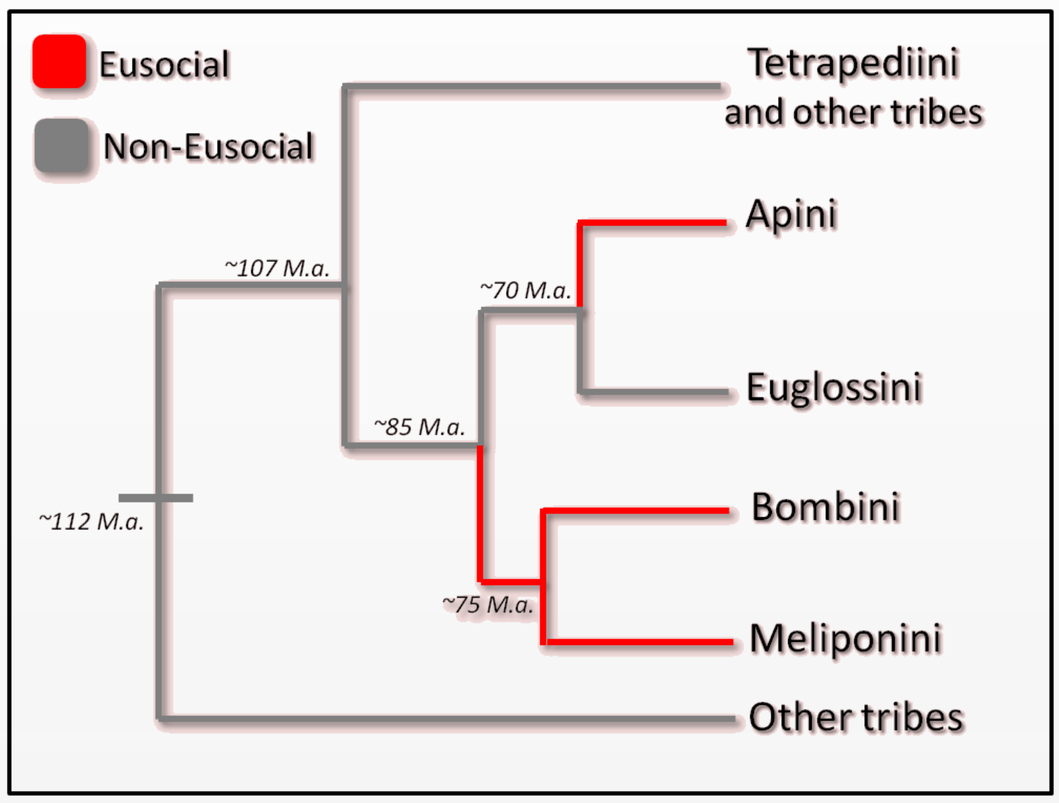

Figure 1: A- *Tetrapedia diversipes*, non-model solitary bee from American tropics; **B-** *Apis mellifera*, widespread eussocial bee used as biological model; **C-** Cladogram of the Apinae subfamily indicating their time of divergence (Based on Cardinal *et al.*, 2010; Fischman *et al.*, 2011; Woodard *et al.*, 2011).

Herein we compare the success of both methods for assembling the transcriptome of the solitary bee *Tetrapedia diversipes* (Figure 1A), a non-model Apinae bee native from the American Tropics. The genome of *Apis mellifera* were used as reference in one of the approaches (Figure 1B), this bee lineage has diverged from *T. diversipes* over 100 million years ago (Figure 1C)

Material and Methods

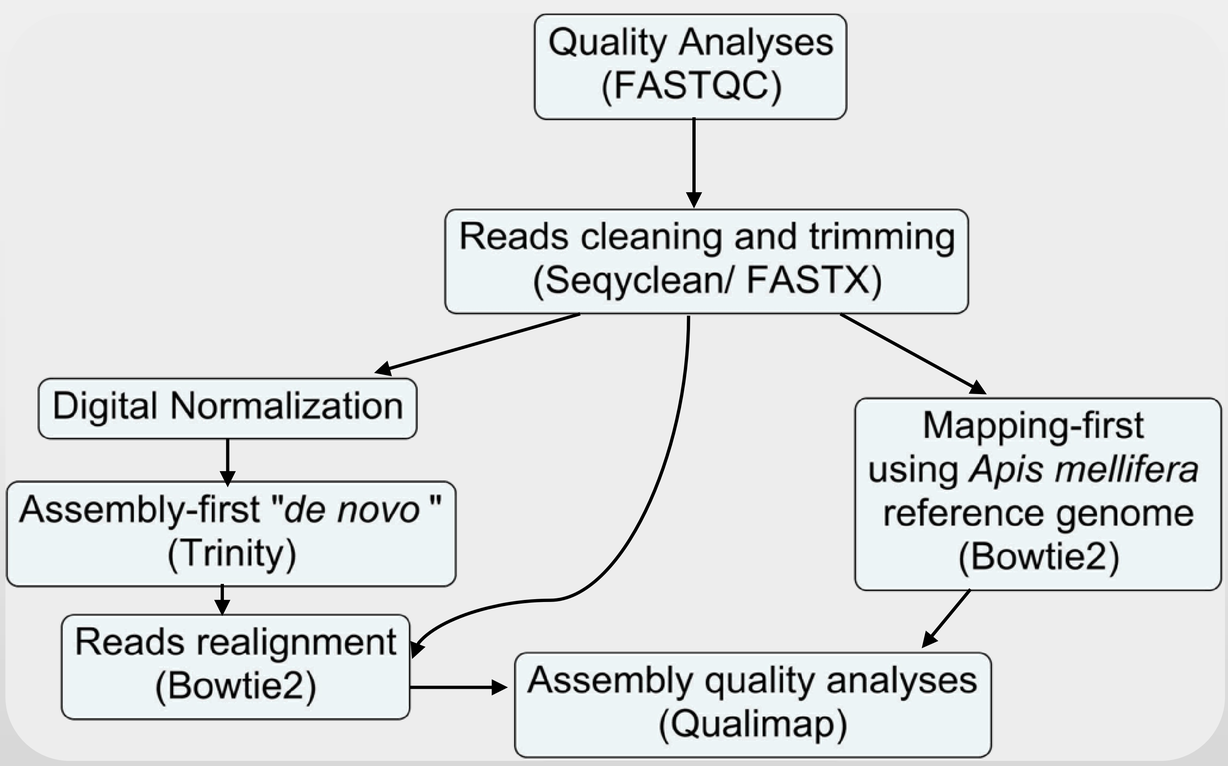

Figure 2: Analyses strategy. In parenthesis follow the programs used in each step.

Results and discussion

Table I: Assembly quality analyses from both methods used.

	Mapping-first	Assembly-first
Reference size	250,270,657	215,715,262
Number of reads	143,522,888	143,522,888
Mapped reads	28,958,837 / 20.18%	141,218,589 / 98.39%
Unmapped reads	114,564,051 / 79.82%	2,304,299 / 1.61%
Mean coverage	7.18	50.17
Coverage Standard Deviation	327.64	977.84
Mean Insert size	61.68	144.93
Median Insert size	132	132

The *de novo* assembly approach seems to be more effective to study non-model species, here *T. diversipes* (Table I). This technique uses more data from the sequenced reads and increases the coverage for each transcript, which improves the following expression analyses.

Acknowledgments: